PadhAI: Information Theory and Cross Entropy

One Fourth Labs

Sigmoid Neuron and Cross Entropy

How does it all tie up to the Sigmoid Neuron

1. Consider the Example:
    a. A signboard with the text **Mumbai**
    b. A random variable X which maps the signboard to: Text, No-Text
    c. The distributions are as follows

X	y (We don't know initially)	ŷ (Predicted using sigmoid)
T	1	0.7
NT	0	0.3

    d. Previously, we were using **Squared-error Loss** = $\Sigma_i (y_i - \hat{y}_i)^2$
    e. Now, we have a better metric, one that is grounded in probability theory (**KL- Divergence**)
    f. KLD(y||ŷ) = $-\Sigma y_i \, log \hat{y}_i + \Sigma y_i \, log y_i$
    g. We aim to minimize loss by KLD with respect to the parameters w, b
    h. From KLD equation, we can see that y_i doesn't depend on w, b. So therefore, we are really only trying to minimize the first term, i.e. the cross-entropy
    i. So in practice, we can treat the second term as a constant, and the equation would really be $min(-\Sigma_i y_i log \hat{y}_i) \quad here\ i \in T, NT$
    j. $Cross\ Entropy\ Loss = \ -1 * log(0.7) - 0 * log(0.3)$
    k. The second term cancels out and we are left with $-log(0.7)$ which is the same as $-log \hat{y} \ \ (for\ the\ true\ case)$
    l. It can be called $-log\ \hat{y}_c$ where c can take the value 0 or 1 which correspond to NT and T